

System Performance Scaling of IBM POWER6™ Based Servers

Jeff Stuecheli

Hot Chips 19 August 2007

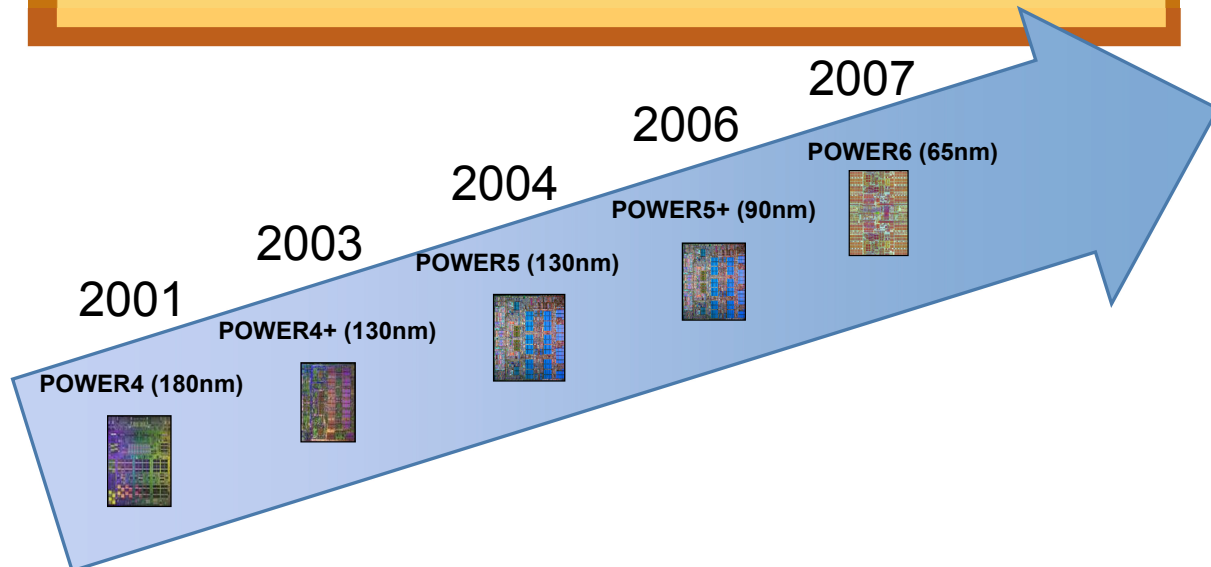
© 2007 IBM Corporation

Agenda

- Historical background
- POWER6™ chip components
- Interconnect topology
- Cache Coherence strategies
- POWER6™ Coherence Mechanism
- System measurements and results

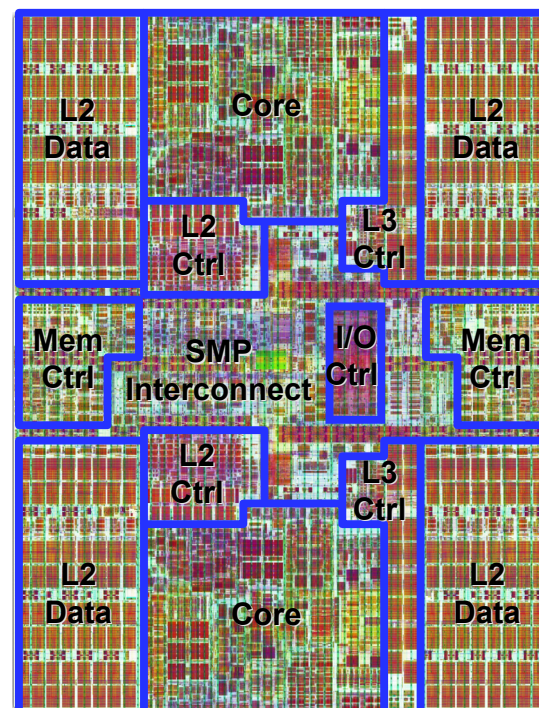
IBM POWER SYSTEMS

Leading up to POWER6™



POWER6™ Chip Overview

- **Ultra-high frequency dual-core chip**
 - 7-way superscalar, 2-way SMT core
 - 9 execution units
 - 2LS, 2FP, 2FX, 1BR, 1VMX, 1DFU
 - 790M transistors
 - Up to 64-core SMP systems
 - 2x4MB on-chip L2
 - 32MB On-chip L3 directory and controller
 - Two memory controllers on-chip
- **Technology**
 - CMOS 65nm lithography, SOI
- **High-speed elastic bus interface at 2:1 freq**
 - I/Os: 1953 signal, 5399 Power/Gnd



POWER6™ Interconnect

Memory DIMMs

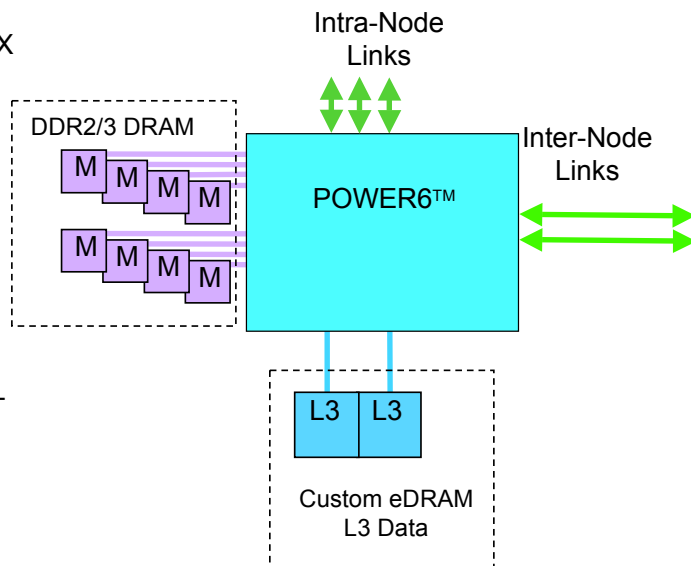
- 8 channels
- Point to Point Elastic interface at 4X DRAM frequency
- Custom on-DIMM buffer chip

L3 data

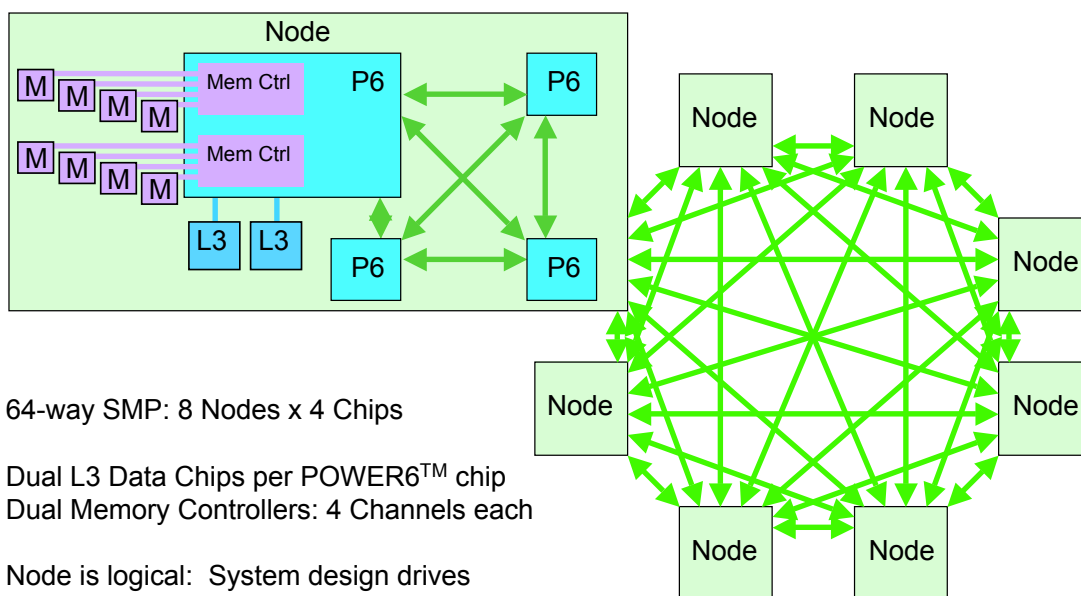
- 32 MB capacity
- Dual 16 byte (8 read + 8 write) interface at ½ CPU frequency

SMP interconnect

- Three Intra-Node SMP buses for 8-way Node
- Two Inter-Node SMP buses for up to 8 Nodes



Example System topology



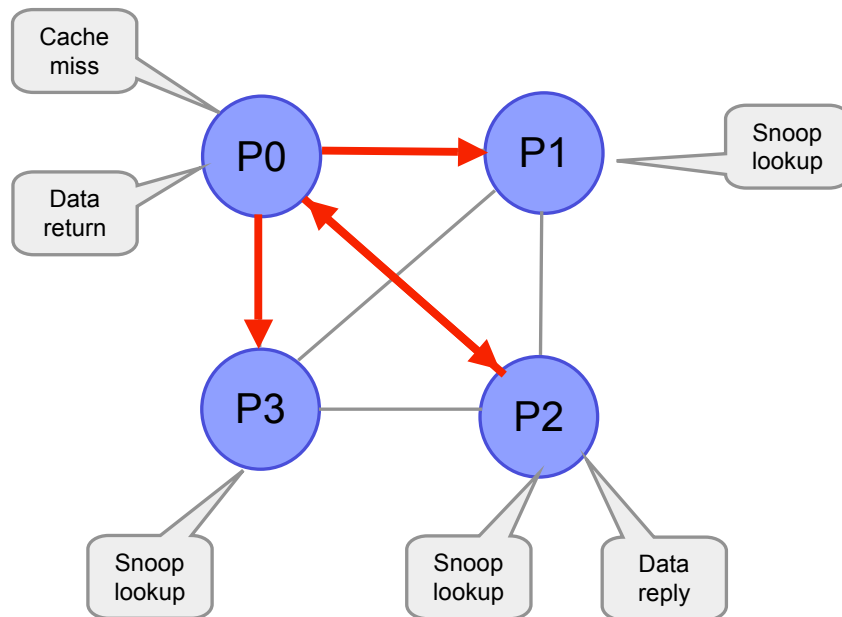
POWER6™ Cache Coherence

- Background
 - Large SMP issues
 - Broadcast and Directory schemes
 - Workload considerations
- POWER6™ scheme details

Cache Coherence strategies

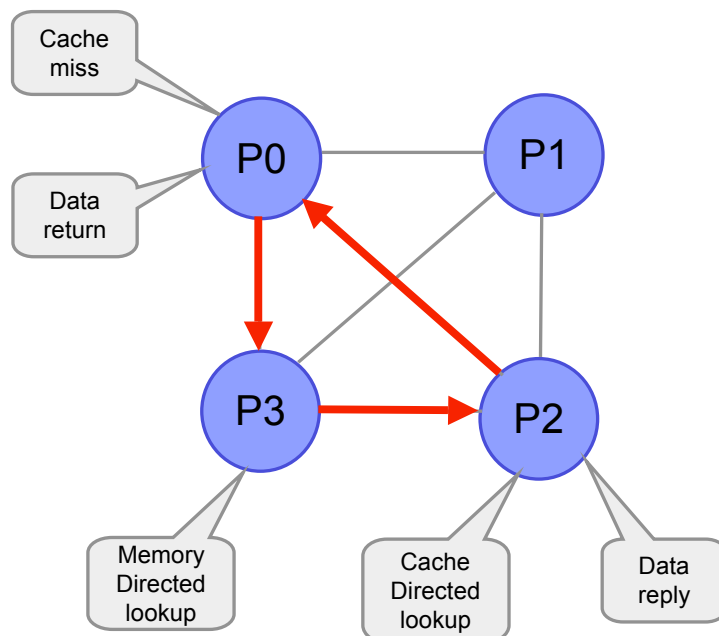
- **Large SMP issues**
 - Data transfer latency
 - Coherence resolution latency
 - Race between DRAM and global snoop
 - Peak throughput
 - Local thread interference

Broadcast Based Coherence



Latency = chip hop + cache read + chip hop

Directory Based Coherence



Latency = chip hop + directory read + chip hop + cache read + chip hop

Broadcast vs Directory Highlights

- **Latency**
 - Memory latency ~equivalent
 - Cache intervention, Broadcast is much faster
- **Request Overhead**
 - Broadcast
 - N directory lookups per request
 - Directory
 - 1 primary directory lookup per request

Workload considerations

- **HPC/Scientific**
 - High bandwidth requirements
 - Regular memory access patterns
- **Commercial/OLTP**
 - High cache 'intervention' rates
 - Low inherent ILP
 - Irregular memory access patterns
- **Virtualization**
 - Workload migration
 - Heterogeneous environment

POWER6™ Approach

■ SubSpace Snooping

- Hybrid approach: Combination of cache snooping and directory lookup
- Design Principle: Provide ideal latency of cache snooping, with high scalability provided with full directories.

■ Design components

- Cache States: MESI enhanced with line location residue.
- Memory Directory: Low cost coherence information stored with cache line data
- Coherence Predictor: Prediction of cache line system state.
- Two Tiered: Node local and System scope

Cache state additions over MESI

■ Invalid State variants

- ***In***: Line invalidated by CPU on same local node
- ***Ig***: Line invalidated by CPU outside local node

■ Valid State variants

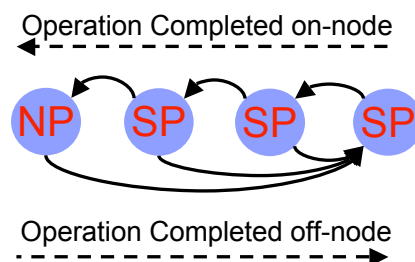
- ***Tn***: Line shared only by CPU on same node. Previously Modified state.
- ***T***: Line shared by CPU outside local node. Previously Modified state.
- ***Ten***: Line shared only by CPU on same node. Previously Exclusive state.
- ***Te***: Line shared by CPU outside local node. Previously Exclusive state.

Memory directory

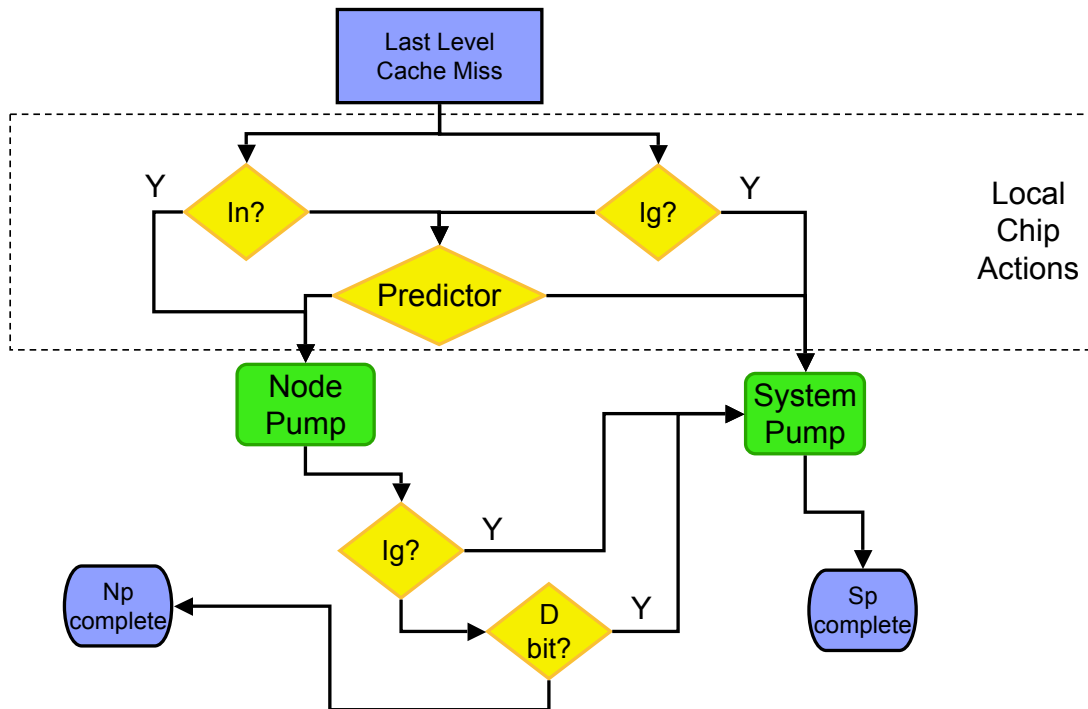
- Single bit per cache line 'buried' in ECC overhead (no additional memory bits added).
- Bit indicates line has moved off node.
- Used as offline coherence check outside critical latency path.

Coherence Predictor details

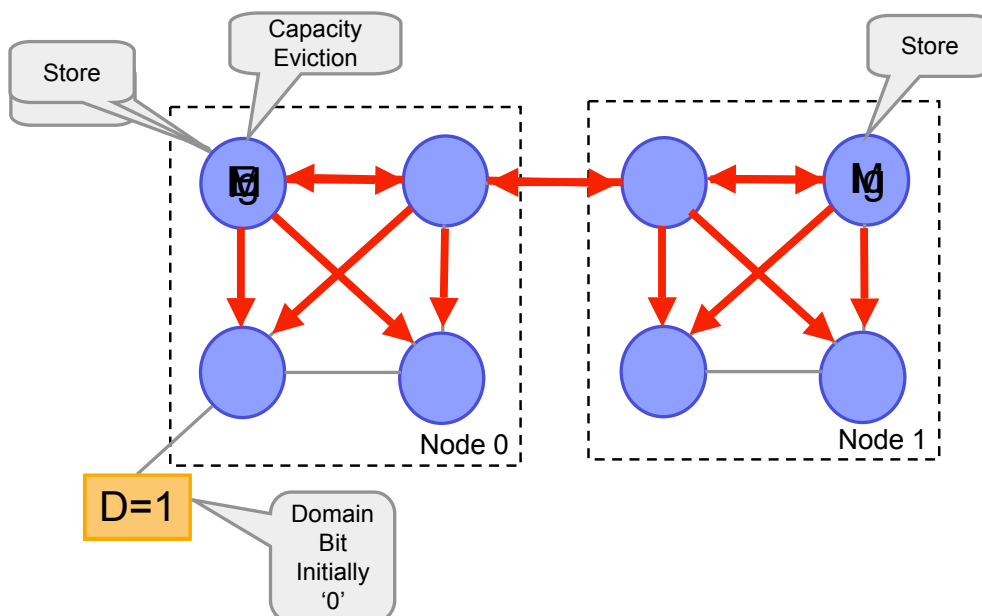
- **Inputs:**
 - System address
 - Thread id
 - Operation type
 - History
- **Predictor:**
 - Array of 2 bit sat counters, biased towards global
 - Local miss-predict penalty greater than Global miss-predict.



Command Broadcast Flow



SubSpace snoop example

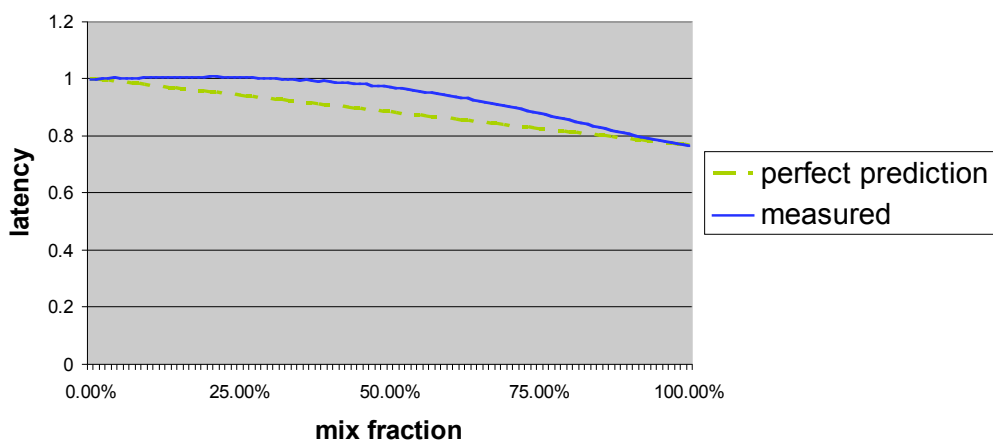


Prediction Accuracy

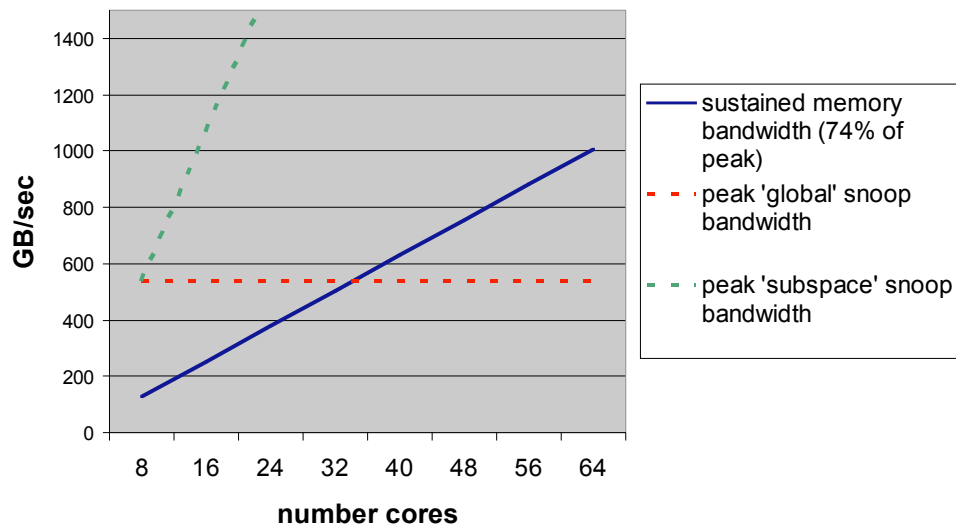
- **Highly accurate cases (~100% correct)**
 - Hot shared lines -> cache residue
 - Locking/synchronizing
 - Embarrassingly parallel -> all local
 - SpecRate
 - Stream
 - Regular Data layout -> consistent regions
 - Regular phases -> temporal predictability
- **Complex data access patterns**
 - Commercial workloads > 95% prediction accuracy

Pathological Latency penalty measurements

- Lmbench based latency measurement
- Mix local and global access in same predictor 'set'
- Testcase forces miss-prediction detection via memory directory bit
- Evaluate worst case latency penalty



Sustained system memory bandwidth running HPC loop



In closing...

- POWER6™ coherence scheme provides latency of broadcast combined with high scalability of distributed directory
- Key Innovation
 - Combining fast predictive broadcast with accurate directory